

Recuperação de informação contextualizada em portais do Governo Federal com base no conteúdo da Lista de Assuntos do Governo

Fabício J. Barth, Antonio Pedro Timoszczuk

¹Diretoria de Inovações e Soluções Tecnológicas
Fundação Aplicações de Tecnologias Críticas - Atech (<http://www.atech.br>)
Rua do Rocio, 313 - 11º andar, Vila Olímpia, São Paulo - SP

{fbarth, antoniop}@atech.br

Resumo. Este artigo propõe a utilização de um sistema de recuperação de informação que expande as consultas do usuário de acordo com o conteúdo da Lista de Assuntos do Governo (LAG). O objetivo deste sistema é permitir uma recuperação de informação mais eficiente nos portais do Governo Federal. Para alcançar este objetivo é proposto: (i) a exploração do uso da estrutura da LAG na apresentação dos documentos recuperados; (ii) a utilização de um WIKI como ferramenta para a edição colaborativa da LAG via Web; e, (iii) um processo para a manutenção da LAG, envolvendo um ciclo de vida das versões com fases para edição colaborativa, aprovação do grupo de trabalho do governo e transferência do conteúdo da LAG para o sistema de recuperação de informação utilizado nos portais.

Abstract. This paper proposes an information retrieval system that user LAG content to expand queries. The system's goal is allow an efficient information retrieval in the Brazil Government's Web portals. To archive this goal, we propose that the information retrieval system: (i) uses the LAG structure to list the documents retrieved, and; (ii) uses a WIKI as a tool for collaborative editing of LAG content. Besides, this work specifies a process for maintaining the LAG content - with collaborative edition and approval steps.

1. Introdução

Os órgãos do Governo Federal disponibilizam milhares de informações e serviços em portais e sítios na Internet, mas o volume e a complexidade da estrutura governamental podem tornar a localização das informações uma tarefa difícil, até mesmo impossível para muitos cidadãos¹.

Para resolver parte desta questão, o Ministério do Planejamento, Orçamento e Gestão, através do projeto e-Ping, iniciou o desenvolvimento da Lista de Assuntos do Governo (LAG), visando facilitar a navegação nos portais do Governo Federal. A LAG é uma lista de indexadores que define conceitos, seus sinônimos e a relação entre os conceitos [ePing 2007]. Uma parte da LAG, com o primeiro e segundo nível de conceitos, é apresentada na Figura 1.

¹Trecho extraído de notícia publicada em <https://www.governoeletronico.gov.br/noticias-e-eventos/noticias/serprogenoticia.2007-04-15.5151017173/?searchterm=LAG>, acessada em junho de 2006

- 1 Agricultura, Extrativismo e Pesca
 - 1.1 Abastecimento
 - 1.2 Agricultura familiar
 - 1.3 Agricultura de subsistência
 - 1.4 Agricultura orgânica
 - 1.5 Alimento
 - 1.6 Assistência técnica
 - 1.7 Defesa Vegetal
 - 1.8 Engenharia agrícola
 - 1.9 Extensão rural
 - 1.10 Extrativismo
 - 1.11 Organização Agrária
 - 1.12 Pecuária
 - 1.13 Pesca
 - 1.14 Política agrícola
 - 1.15 Produção vegetal
 - 1.16 Produto vegetal
 - 1.17 Solos
 - 1.18 Transgenia e seqüenciamento
- 2 Ciência, Informação e Comunicação
 - 2.1 Ciência e tecnologia
 - 2.2 Comunicação
 - 2.3 Informação: Gestão, Preservação e Acesso
- 3 Comércio, Serviços e Turismo
 - 3.1 Comércio e Serviços
 - 3.2 Turismo
- 4 Cultura, Lazer e Esporte
 - 4.1 Cultura
 - 4.2 Esporte
 - 4.3 Lazer

Figura 1. Uma parte da Lista de Assuntos do Governo, com o primeiro e segundo nível de conceitos

Segundo o projeto e-Ping, essa lista vai possibilitar que os cidadãos possam encontrar informações nos portais do Governo Federal independentemente do conhecimento que possuem acerca da estrutura formal do Governo [ePing 2007]. Para tanto, é necessário que a estrutura das informações existentes nos portais do Governo Federal esteja organizada de acordo com a estrutura da LAG ou os sistemas de recuperação de informação existentes nos portais façam uso da informação contida na LAG para recuperar os documentos de interesse dos usuários.

Na tentativa de solucionar os problemas relacionados com os sistemas de recuperação de informação, inúmeros trabalhos têm proposto métodos alternativos para aumentar a precisão e o índice de recuperação de documentos relevantes para o usuário. Alguns destes trabalhos propõem o uso de ontologias para expandir a consulta do usuário. Ao permitir a expansão da consulta do usuário a partir de termos adicionais próprios do domínio, a ontologia viabiliza a recuperação de documentos que seriam ignorados pela consulta original melhorando, por conseguinte, o resultado final para o usuário [Barth and Timoszczuk 2008, Bonino et al. 2004, Bhogal et al. 2007].

Levando-se em consideração as iniciativas do Governo Federal e as soluções existentes para recuperação de informação que utilizam ontologias e taxonomias, este texto propõe:

- A utilização de um sistema de recuperação de informação que expande as consultas do usuário de acordo com o conteúdo da Lista de Assuntos do Governo (LAG), visando uma recuperação de informação mais eficiente nos portais do Go-

verno Federal.

- A apresentação dos resultados recuperados utilizando a própria estrutura da Lista de Assuntos do Governo (LAG).
- A especificação do suporte computacional necessário para que qualquer cidadão possa, de maneira colaborativa, editar a Lista de Assuntos do Governo (LAG) via Web.
- A definição de um processo para a manutenção da Lista de Assuntos do Governo (LAG), envolvendo um ciclo de vida das versões com fases para edição colaborativa, aprovação do grupo de trabalho do governo e transferência do conteúdo da LAG para o sistema de recuperação de informação.

Este texto apresenta uma proposta que une diversos conceitos e tecnologias, aparentemente não relacionados, para resolver um problema de recuperação de informação em portais do governo. Neste texto não é apresentada nenhuma validação da proposta completa, exceto, referências para outros textos que descrevem os resultados alcançados com alguns componentes que fazem parte da solução. Sendo assim, este texto está estruturado da seguinte maneira: na seção 2 é apresentada a proposta e na seção 3 são apresentadas algumas considerações sobre a proposta.

2. Proposta

Em [Barth et al. 2007] é descrito um sistema de recuperação de informação projetado para processar fontes de informação estruturada e não-estruturada, sejam elas públicas ou privadas. Tais fontes são pesquisadas a partir de uma consulta onde os termos escolhidos pelo usuário podem ser expandidos mediante confronto com uma ontologia de domínio. Os documentos recuperados são então submetidos a algoritmos de agrupamento e de identificação de entidades nomeadas (pessoas, organizações, locais e termos importantes para o domínio da aplicação). Ambas são técnicas de mineração de texto que colocam em evidência as relações entre os documentos e permitem a apresentação gráfica das mesmas ao usuário.

Ao permitir a expansão da consulta do usuário a partir de termos adicionais próprios do domínio, a ontologia viabiliza a recuperação de documentos que seriam ignorados pela consulta original. Um exemplo é apresentado na Figura 2. Neste exemplo, a consulta “lazer” (inserida no campo texto) é substituída, através de um processo transparente para o usuário, pela consulta “lazer cultura esporte” - permitindo a recuperação de documentos que estão relacionados com o tema “lazer”, mas que não possuem a palavra “lazer”. Na Figura 2 as palavras cultura, lazer e esporte que aparecem nos documentos recuperados estão destacadas.

No contexto da engenharia do conhecimento, uma ontologia é especificada sob a forma de um vocabulário que representa os conceitos do domínio [Uschold and Gruninger 1996]. Um exemplo simples, que corresponde à ontologia utilizada pelo sistema, é o da hierarquia de tipos, onde são especificadas classes (conceitos) e seus relacionamentos com superclasses e subclasses.

Esse tipo de ontologia resulta, portanto, da decomposição do domínio em conceitos que se relacionam com outros mais genéricos e mais específicos. Todos os conceitos podem ter sinônimos associados. Os sinônimos permitem a cobertura exhaustiva do voca-



Figura 2. Exemplos de documentos recuperados com a consulta “lazer”

bulário do domínio, enquanto que a hierarquia de tipos permite a navegação de conceitos mais específicos para mais genéricos, e vice-versa (Figura 3).

A consulta submetida pelo usuário é contextualizada da seguinte maneira: cada termo que compõe a consulta é pré-processado tendo em vista a remoção de acentos, de maiúsculas e a redução ao singular. Em seguida, o termo é confrontado com aqueles definidos na ontologia e seus sinônimos. Se o termo ou um de seus sinônimos for encontrado na ontologia, todos são conectados com o operador *OR*, compondo uma expressão que vai substituir o termo original na consulta.

Caso o termo não esteja na ontologia, ele é mantido sem alterações na consulta. Todos os termos, expandidos ou não, são conectados com o operador *AND* para formar a consulta contextualizada. Termos compostos podem fazer parte tanto da consulta como da ontologia. Tais termos, que devem ser especificados entre aspas pelo usuário, não são reduzidos ao singular.

A critério do usuário, a consulta original ou contextualizada pode ser refinada mediante duas operações: focalização e generalização. Na generalização, os termos da consulta e seus sinônimos são substituídos pelos termos correspondentes à(s) superclasse(s) e seus sinônimos. Na focalização, são acrescentados aos termos da consulta e seus sinônimos os termos correspondentes à(s) subclasses(s) e seus sinônimos.

O objetivo da primeira operação é obter uma consulta mais abstrata (composta de termos mais genéricos na hierarquia); o da segunda, obter uma consulta mais específica (composta de termos mais específicos na hierarquia). Isto permite ao usuário navegar a ontologia tendo em vista o ajuste de sua consulta ao nível de abstração desejado [Barth and Timoszczuk 2008].

Os testes e experimentos realizados, utilizando uma metodologia de avaliação ade-

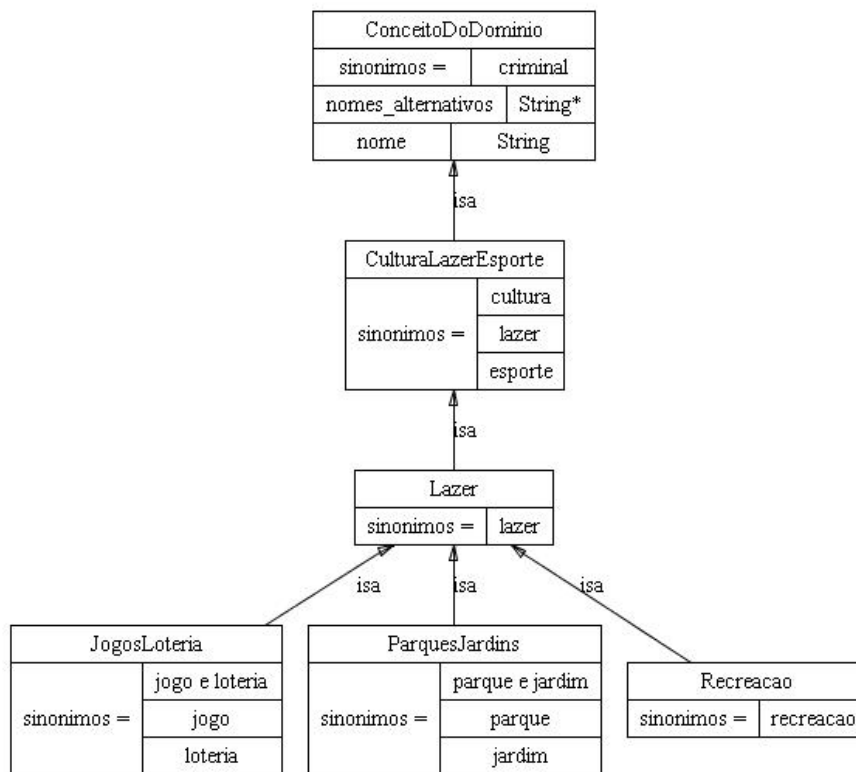


Figura 3. Trecho da ontologia com os conceitos relacionados com “lazer”

quada para sistemas de recuperação de informação, demonstraram que as operações que fazem uso da ontologia (contextualizar, focalizar e generalizar) possuem um desempenho superior quando comparadas a um processo de recuperação de informação convencional [Barth and Timoszczuk 2008].

2.1. Apresentação dos documentos recuperados

Tipicamente, o resultado de uma consulta em um sistema de recuperação de informação é uma lista de objetos [Danilowicz and Nguyen 2002]. Essas listas de objetos recuperados são usualmente extensas, o que dificulta o processo de exploração dos resultados. Essa forma de apresentação não tem uma preocupação com a eficiência e, muito menos, com o nível de satisfação do usuário [Sheiderman 1997]. A satisfação do usuário é maior quando o mesmo utiliza uma interface homem máquina que lhe permita realizar a atividade de recuperação de informação de maneira mais imediata e com uma linguagem que lhe seja familiar. Este trabalho propõe a exploração do uso da estrutura da ontologia ou taxonomia na apresentação dos documentos recuperados. Tem-se como hipótese, que ao agrupar os documentos recuperados em itens da ontologia, com termos familiares ao usuário, a navegação pelos resultados será mais eficiente do que através de uma lista extensa de documentos.

Na Figura 4 é possível visualizar um exemplo da utilização da estrutura da LAG para a apresentação dos documentos recuperados. Cada caixa corresponde a um conceito existente na LAG. Os documentos recuperados que pertencem a um conceito devem ser apresentados dentro desta caixa.

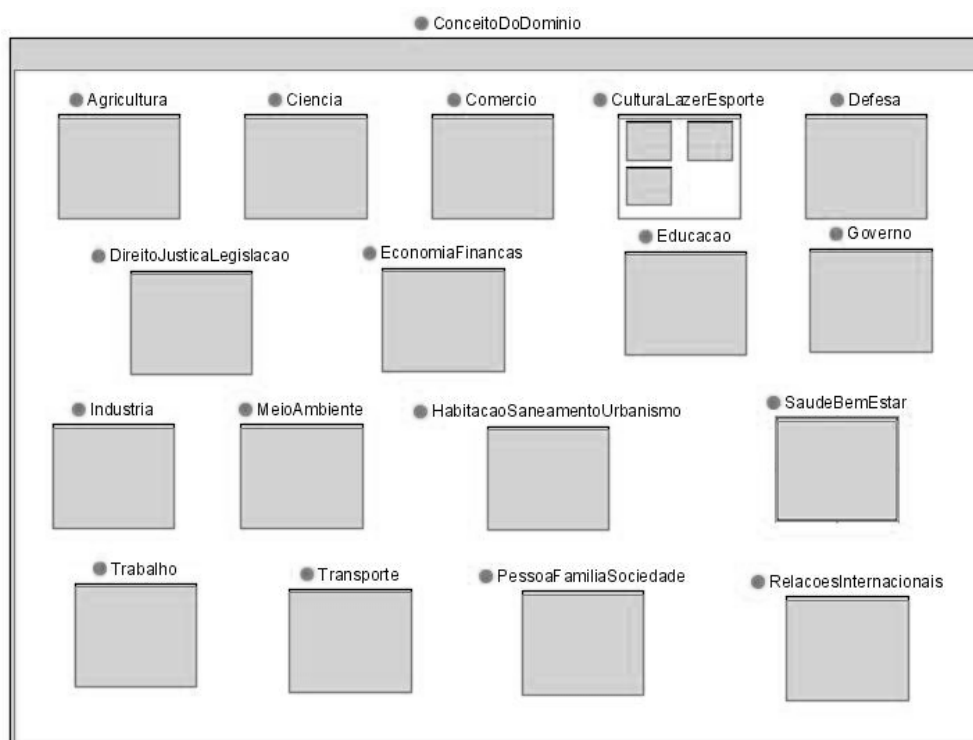


Figura 4. Apresentação dos documentos recuperados usando a estrutura da ontologia ou LAG

Outras formas para apresentação dos resultados, que utilizam a estrutura da LAG, podem ser explorados. Por exemplo, documentos recuperados inseridos dentro de uma árvore no formato da LAG.

2.2. Utilização e construção colaborativa da Lista de Assuntos do Governo (LAG)

A Lista de Assuntos do Governo (LAG) é facilmente traduzida para a ontologia utilizada pelo sistema descrito em [Barth et al. 2007]. Ao utilizar o conteúdo da LAG, na ontologia de domínio, sobre uma base de dados com documentos do Governo Federal, pode-se garantir uma recuperação de informação mais eficiente nos portais do Governo Brasileiro. No entanto, é fácil perceber que o desenvolvimento de uma ontologia ou lista de assuntos é uma tarefa laboriosa. Uma ontologia completa é aquela que possui todos os principais conceitos do domínio. Além disso, a equipe que desenvolve a ontologia pode definir conceitos e sinônimos não compartilhados pelos usuários do sistema de recuperação de informação. Se isto acontecer, os resultados retornados pelo sistema de recuperação de informação não serão os resultados de interesse do usuário.

Percebe-se que o Governo Brasileiro tem a mesma preocupação quanto ao conteúdo da Lista de Assuntos do Governo (LAG), pois o mesmo disponibiliza a lista para consulta pública, permitindo que qualquer cidadão possa contribuir para o conteúdo da lista:

“Para ouvir a opinião da sociedade a respeito, a Secretaria de Logística e Tecnologia da Informação (SLTI) disponibilizou a LAG para

consulta pública neste portal (consultas públicas - consultas em andamento).”²

Tem-se como hipótese que ao transformar a tarefa de construção da ontologia em uma tarefa colaborativa consegue-se: (i) reduzir o problema de interpretação durante a construção de uma ontologia, pois os conceitos utilizados na ontologia e na LAG serão fornecidos pelos próprios usuários, e; (ii) tornar a tarefa de construção de ontologias menos laboriosa.

Para a construção colaborativa de ontologias ou taxonomias propõe-se o uso de um WIKI. O termo WIKI é utilizado para identificar um tipo específico de coleção de documentos em hipertexto ou o software colaborativo usado para criá-lo. Um WIKI permite que os documentos sejam editados colaborativamente através da utilização de um navegador Web, permitindo a geração e distribuição rápida de conteúdo e a fácil colaboração entre usuários. Tipicamente, um WIKI permite que usuários possam: (i) adicionar novo conteúdo; (ii) ligar um conteúdo ao outro através de hiperlinks; (iii) editar conteúdo já existente; (iv) organizar e estruturar o conteúdo; (v) visualizar o conteúdo, e; (vi) acessar o histórico de contribuições. Adicionalmente, Wikis usam vários mecanismos para monitorar o histórico de contribuições dos usuários [Wikipedia 2006].

As vantagens do uso de WIKI, como ferramenta para edição colaborativa de ontologias, são: (i) Wikis geram uma rede de conhecimento ligando pessoas ao conteúdo criado; (ii) Wikis auxiliam na criação de documentos consensuais, e; (iii) através de Wikis é fácil gerar informações a partir de diversas fontes [O’Leary 2008, Barth et al. 2008].

Nos últimos anos, inúmeras organizações vêm explorando Wikis como ferramentas para: Gestão de Projetos [Xiao et al. 2007, Louridas 2006], Gestão de Conhecimento [Hasan and Pfaff 2006], Levantamento de Requisitos junto ao cliente [Wagner and Majchrzak 06 7], Levantamento de Idéias entre os colaboradores [Majchrzak et al. 2006], entre outras atividades. Entre as organizações que utilizam este tipo de ferramental estão: IBM³, NASA⁴, Google e Microsoft.

As soluções de Wikis disponíveis no mercado são soluções robustas, escaláveis e que adotam os padrões de segurança do mercado. Por exemplo, em [Barth et al. 2008] é descrito a utilização da plataforma XWIKI⁵ como ferramenta para auxiliar o processo de inovação de uma organização. Atualmente, neste WIKI existem 41 espaços, 914 páginas e 233 usuários. Inicialmente, este WIKI foi criado apenas para suprir as necessidades da área de Inovação. No entanto, com o passar do tempo, projetos em execução e comunidades de prática começaram a utilizar o ambiente colaborativo também. Sendo assim, acredita-se que a utilização de um WIKI como ferramenta para a edição colaborativa de ontologias e taxonomias via Web é uma abordagem viável. Claro que o número de usuários neste caso seria maior do que o caso apresentado em [Barth et al. 2008].

²<https://www.governoeletronico.gov.br/noticias-e-eventos/noticias/serprogenoticia.2007-04-15.5151017173/?searchterm=LAG>

³Exemplo de utilização de Wikis para discussão sobre a tecnologia LOTUS da IBM: <http://www-10.lotus.com/ldd/lcwiki.nsf>

⁴Exemplo de Wiki utilizado por grupos de trabalho da NASA: <http://wiki.nasa.gov>

⁵<http://www.xwiki.org>

2.3. Processo para edição, manutenção e utilização do conteúdo da LAG

Para alcançar os quatro objetivos enumerados no início deste texto é proposto um processo para edição colaborativa e manutenção da Lista de Assuntos do Governo. Este processo é composto pelas seguintes etapas:

1. Os usuários (cidadãos) podem alterar livremente o conteúdo da LAG em um ambiente colaborativo disponibilizado em um endereço WEB.
2. Periodicamente, a última versão da LAG é disponibilizada para o grupo de trabalho do governo.
3. O grupo de trabalho do governo realiza a validação do conteúdo da LAG, autorizando a sua publicação.
4. A nova versão da LAG, autorizada para publicação, é disponibilizada no ambiente colaborativo.
5. A nova versão da LAG, autorizada para publicação, é traduzida para o formato utilizado pelo sistema de recuperação de informação contextualizada.

A Figura 5 ilustra a interação entre os atores, as etapas e os componentes que fazem parte do processo. Os números destacados na Figura correspondem as etapas do processo definidas acima.

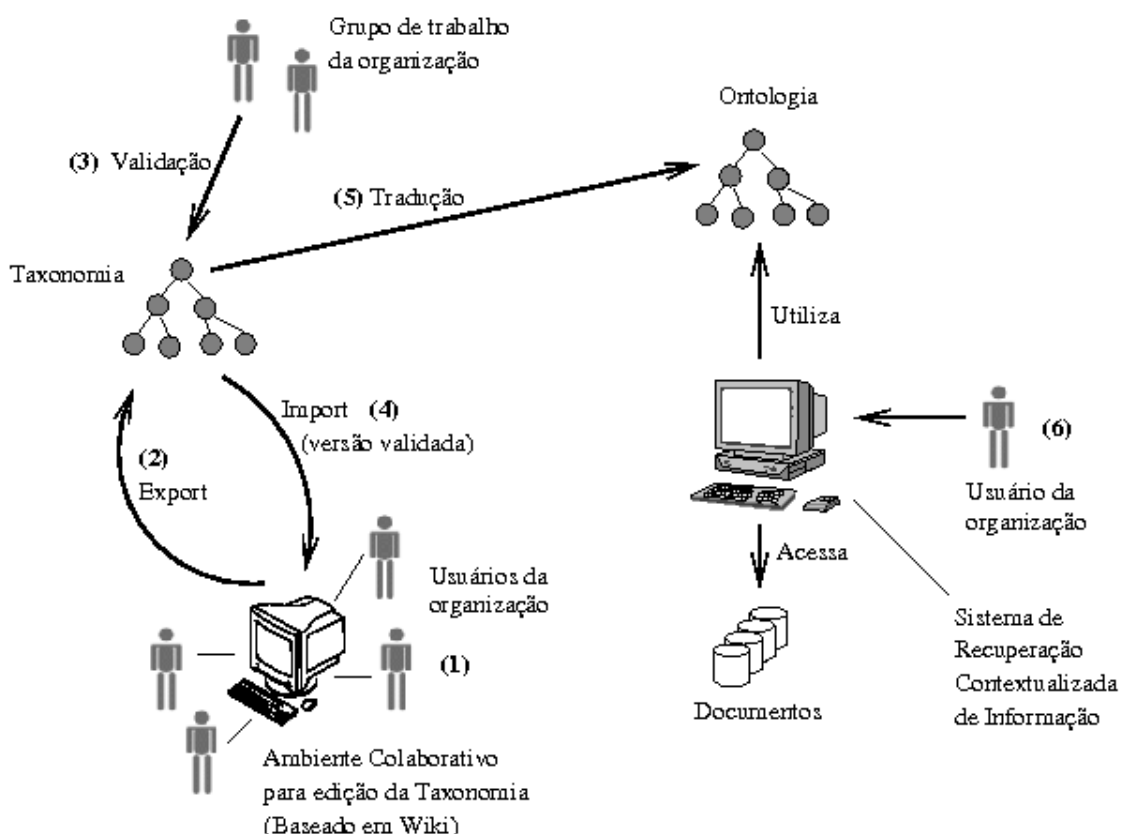


Figura 5. Processo para edição, manutenção e utilização do conteúdo da LAG

Em paralelo ao processo descrito acima, os usuários podem utilizar o sistema de recuperação de informação contextualizada a qualquer momento nos portais do governo federal.

3. Considerações Finais

Este texto propôs a utilização de um sistema de recuperação de informação que expande as consultas do usuário de acordo com o conteúdo da Lista de Assuntos do Governo (LAG), visando uma recuperação de informação mais eficiente nos portais do Governo Federal. Além disso, foram propostos: (i) a exploração do uso da estrutura da LAG na apresentação dos documentos recuperados; (ii) a utilização de um WIKI como ferramenta para a edição colaborativa da LAG via Web; e, (iii) um processo para a manutenção da LAG, envolvendo um ciclo de vida das versões com fases para edição colaborativa, aprovação do grupo de trabalho do governo e transferência do conteúdo da LAG para o sistema de recuperação de informação.

Parte do que foi proposto neste trabalho ainda não foi validado. Por exemplo, para avaliar se o processo de desenvolvimento colaborativo de ontologias elimina ou reduz o problema de interpretação e torna a construção de ontologias menos laboriosa será necessário executar uma avaliação qualitativa observacional [Yin 2005] na forma de um estudo de caso. O mesmo acontece com a avaliação da interface homem-máquina. Durante a avaliação da interface homem-máquina deverão ser considerados os seguintes critérios: efetividade da interface, eficiência da interface e satisfação dos usuários. Além disso, é importante avaliar a segurança e escalabilidade de toda a solução.

Referências

- Barth, F. J., Belderrain, M. C. R., Quadros, N. L. P., Ferreira, L. L., and Timoszczuk, A. P. (2007). Recuperação e mineração de informações para a área criminal. In *ENIA VI - Encontro Nacional de Inteligência Artificial. Anais do XXVII Congresso da SBC*, pages 1292–1301.
- Barth, F. J. and Timoszczuk, A. P. (2008). Expansão automática de consultas utilizando ontologias. In *Seminário de Pesquisa em Ontologia no Brasil*, pages 1–8.
- Barth, F. J., Vilela, G. T., Timoszczuk, A. P., and Mussoi, A. L. G. (2008). Wiki as a tool for improving the innovation process. In *Proceedings of the 4th International Symposium on Wikis*, pages 1–2.
- Bhokal, J., Macfarlane, A., and Smith, P. (2007). A review of ontology based query expansion. *Inf. Process. Manage.*, 43(4):866–886.
- Bonino, D., Corno, F., Farinetti, L., and Bosca, A. (2004). Ontology driven semantic search. *WSEAS Transaction on Information Science and Application*, 1(6):1597–1605.
- Danilowicz, C. and Nguyen, H. C. (2002). Using user profiles in intelligent information retrieval. In Hacid, M.-S., Rás, Z. W., Zighed, D. A., and Kodratoff, Y., editors, *Foundations of Intelligent Systems. 13th International Symposium*, LNAI 2366, pages 223–231, Lyon, France. Springer-Verlag.
- ePing (2007). Lista de assuntos do governo: Taxonomia para navegação. Technical report, e-Ping: Padrões de Interoperabilidade de Governo Eletrônico.
- Hasan, H. and Pfaff, C. C. (2006). The wiki: an environment to revolutionise employees' interaction with corporate knowledge. In *OZCHI '06: Proceedings of the 18th Australia conference on Computer-Human Interaction*, pages 377–380, New York, NY, USA. ACM.

- Louridas, P. (2006). Using wikis in software development. *IEEE Softw.*, 23(2):88–91.
- Majchrzak, A., Wagner, C., and Yates, D. (2006). Corporate wiki users: results of a survey. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 99–104, New York, NY, USA. ACM.
- O'Leary, D. E. (2008). Wikis: From each according to his knowledge. *IEEE Computer Society*, pages 34–41.
- Shneiderman, B. (1997). *Designing the User Interface: Strategies for Effective Human-computer Interaction*. Addison-Wesley, Reading, MA.
- Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*.
- Wagner, C. and Majchrzak, A. (06-7). Enabling customer-centricity using wikis and the wiki way. *J. Manage. Inf. Syst.*, 23(3):17–43.
- Wikipedia (2006). Wiki. Adquirido no site <http://en.wikipedia.org/wiki/Wiki> em junho de 2008.
- Xiao, W., Chi, C., and Yang, M. (2007). On-line collaborative software development via wiki. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 177–183, New York, NY, USA. ACM.
- Yin, R. K. (2005). *Estudo de caso: Planejamento e Métodos*. Bookman, 3ª edição edition.